1-2002

An examination of the accuracy of slogans related to assessment centres

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg
DOI: https://doi.org/10.1108/00483480210412436

An examination of the accuracy of slogans related to assessment centres

Filip Lievens (Ghent University, Ghent, Belgium)

Published in Personnel Review, Volume 31, Issue 1-2, 2002, Pages 86-102

https://doi.org/10.1108/00483480210412436

Abstract:

To diffuse new management fads or to promote existing management practices slogans are often used. In this study eight slogans regarding assessment centres were retrieved from brochures of consultancy agencies and were compared to empirical research. These slogans were that assessment centres work, are expensive, fair, have a potent influence on candidates, provide insight into candidates' personality, and are not prone to trainability effects. Other slogans include that trained assessors speak the same language and that candidate reactions are positive. Resulting from this comparison a more fine-grained picture of the strengths and weaknesses of assessment centres emerges. In particular, it is demonstrated that the accuracy of the slogans often depends on design factors. It is suggested that future studies combine multiple research perspectives, examine assessment centre process issues, and validate new assessment centre applications.

Keywords: Assessment centres, Validity, Reliability, Training, Perception

Introduction

In advertising campaigns catchy slogans are often used to persuade consumers to purchase a particular product or brand. The same is true for political campaigns where slogans aim to succinctly convey the main message of a politician or party. However, the use of slogans is not restricted to advertising and politics as discourse, slogans, and common beliefs also play an important role to diffuse new management fads or to promote existing management practices (Abrahamson, 1996; Johns, 1993).

In this study we delve deeper into slogans related to a particular management technique, namely assessment centres (ACs). Although ACs were first applied in the military, they have been increasingly used over the last 30 years in both private and public organisations. Recent surveys show that ACs are still gaining in popularity with estimates of their use in the UK, for example, varying between 44 per cent and 60 per cent (Keenan, 1995; Shackleton and Newell, 1994). The aim of this article consists in comparing the slogans, which consultancy agencies use to promote ACs, to empirical research evidence. Accordingly, we aim to present a fine-grained and empirically-based portrayal of the strengths and weaknesses of ACs.

It is important to test the accuracy of slogans in general and AC slogans in particular because slogans may lead organisations to use management techniques such as ACs without a thorough understanding of their strengths and weaknesses. This is because slogans may typically present a simplified and unrealistic portrayal of ACs. This simplified view may lead organisations to cut down on essential AC components such as the number of exercises, the use of psychologist assessors, the assessor training, or the integrative assessor discussion (Boyle et al., 1995; Lievens and Goemaere, 1999). However, many of these

components have been shown to improve the predictive validity of ACs (Gaugler et al., 1987). In other words, when organisations rely only on slogans and do not consider the empirical research evidence, this may lead to quick-fix approaches and poor AC design. As a consequence, ACs may eventually become the victims of their own popularity.

Method

The method used in this study consisted of two phases. In the first phase we searched for slogans related to ACs. The following three criteria for inclusion were used. First, the slogans had to be presented in commercial brochures of consultancy agencies. In this study these were Flemish and Dutch consultancy agencies. Second, these slogans had to be used to promote the practice of ACs to companies. Third, the slogans had to describe the strengths or weaknesses of ACs in one short sentence. For example, slogans were often presented in the margin by way of a summary of the main text.

Eight nonredundant slogans were found. The remainder briefly explains each of the following slogans and confronts them with the available empirical research:

- "Trained assessors speak the same language".
- "ACs are expensive".
- "ACs are fair".
- "Candidates react enthusiastically to ACs".
- "ACs have a potent influence on candidates".
- "As a candidate it is difficult to prepare oneself for an AC".
- "ACs provide insight into candidates' personality".

In the second phase we searched for empirical research studies to (in)validate the slogans. To this end, we conducted a search using a number of computerized databases (i.e. PsychLit, Current Contents, Dissertation Abstracts International, and the Social Science Citation Index). We used key words such as "assessment centre", "validity", "fairness", "reliability", "trainability", "applicant reactions". Although we looked especially for recent studies, no specific publication year range was set. We also scrutinized reference lists from obtained studies to find other published and unpublished studies. At last, researchers in the assessment centre domain were contacted to retrieve more unpublished papers.

Overview of slogans vs empirical evidence

Trained assessors speak the same language

This slogan posits that assessors agree to a large extent regarding the ratings made of AC candidates. However, when we compare this statement to empirical results, a more fine-grained picture emerges. In fact, when different assessors are asked to evaluate a candidate on a given dimension within an exercise (the so-called within-exercise dimension ratings), they tend to agree only moderately. Jones (1981), for instance, reported inter-rater reliability coefficients for within-exercise dimension ratings from 0.38 to 0.67. In other studies inter-rater reliability equalled 0.50 (Tziner and Dolan, 1982) or 0.53 (Tziner et al., 1993). Chan (1996), however, reported higher values (0.84).

Assessors' agreement increases when required to make global summary judgements about assessees. For example, Jones (1981) reported inter-rater reliability coefficients from 0.65 to 0.73 for overall exercise performance ratings. Similar increases in inter-rater reliability have been reported by Borman (1982) and Tziner and Dolan (1982). In a more recent study inter-rater reliability for overall exercise ratings ranged from 0.69 to 0.99 with a mean of 0.93 (Gatewood et al., 1990).

In addition, assessors seem to agree to a higher degree when they have the opportunity to discuss candidates' performance with one another prior to providing ratings. Jones (1981) found that under this condition the inter-rater reliability of overall exercise performance ratings rose from 0.73 to 0.83.

Another factor, which influences inter-rater reliability, is the level of standardisation of an AC exercise. In this regard Gatewood et al. (1990) correlated overall ratings of the same participants in two different leaderless group discussions. Disappointingly, these correlations varied between 0.35 and 0.62, suggesting that the unstructured nature of leaderless group discussions seriously affected reliable scoring.

The number of assessors seems to be a fourth factor influencing inter-rater reliability. Lievens (in press) examined the effects of reducing the number of assessors. Diminishing the number of assessors from 3 to 1 had a serious impact on the reliability coefficient as it dropped from about 0.80 to 0.60.

A fifth and last important factor is the type of training provided to assessors. Lievens (2001) found higher inter-rater reliabilities for assessors, whom training provided with the same evaluative standards as a frame-of-reference for judging performance. These assessors outperformed assessors receiving behaviour observation training or no training.

In sum, reliability studies in the AC field have mainly examined the inter-rater reliability of within-exercise ratings. Generally, moderate values were found, lending partial support for the slogan mentioned at the outset. However, more importantly in recent years various factors, which may lead assessors "to speak a similar language", have been identified. These factors include the amount of behavioural evidence available, the use of an integrative discussion, the level of standardisation of AC exercises, the number of assessors, and the type of assessor training. Future studies are needed to identify other factors, which lead to more reliable assessor ratings. These factors might include the type of assessor, the use of behavioural checklists, or the integrative discussion format.

Assessment centres work: they show high predictive validity

This second slogan can best be answered by looking at the meta-analysis of AC criterion-related validity studies (Gaugler et al., 1987). This meta-analysis contained 107 validity coefficients from 50 independent studies, which were conducted between 1950 and 1984. The mean corrected validity coefficient equalled 0.37, endorsing the slogan of the strong predictive power of ACs. Examination of possible moderator variables revealed that ACs were more predictive when they consisted of a larger number of situational exercises, when managers and psychologists served as assessors, and when peer ratings were used. Further, higher validities were obtained for job potential ($r = 0.53$) than for job performance ($r = 0.36$).

During the last decade, the search for other moderators of AC criterion-related validity has remained an important research theme. For example, Moser et al. (1999) studied the moderating effect of assessor-assessee acquaintance on the predictive validity of ACs. When assessor-assessee acquaintance was less than or equal to two years, the validity of the AC was 0.09. This value increased dramatically to 0.50 when assessor-assessee acquaintance was greater than two years. Jones et al. (1991) attempted to improve the predictive validity of an AC by introducing procedural changes (e.g. limiting the number of

dimensions to be rated, changing the scoring system, etc.). However, no significant improvement in predictive validity was obtained. Bartels and Doverspike (1997) investigated whether differences in level (i.e. upper and middle) and business stream (i.e. chemical, corporate, distributions, and research) moderated AC validities. They concluded that these validities did not increase when disaggregated according to either level or business stream.

Taken together, the message emerging from this body of research is twofold. First, meta-analytic research confirms that ACs work and shows that AC ratings are predictive of a variety of criterion measures, supporting this second slogan about ACs. Second, practitioners should realise that the viability of this slogan depends on the AC under investigation because various moderators (e.g. number of exercises, type of assessor, the degree of assessee-assessor acquaintance) have been discovered.

Assessment centres are expensive

One of the widespread criticisms of ACs is related to their high costs. For example, the AC is often referred to as "the Rolls-Royce of selection methods" (Cook, 1998, p. 191). Utility studies can put this statement to the test, as they examine the degree to which a selection procedure improves the quality of applicants chosen over random selection. In general, the data necessary to determine the utility of a selection procedure include the validity of the procedure, the selection ratio, and the criterion standard deviation in dollars of job performance. Estimates of the centre design and administration costs have also to be available.

The preponderance of previous utility studies of ACs have been conducted in the United States. Cascio and Silbey (1979) discovered that even ACs with validities as low as 0.10 possess positive gains in utility over random selection. Other earlier studies demonstrated that ACs had more utility than unstructured interviews (Burke and Frederick, 1986; Cascio and Ramos, 1986) and personality inventories (Hogan and Zenke, 1986).

The last decade is marked by a paucity of studies on AC utility. Goldsmith (1990) compared three procedures to select first-line supervisors: a paper-and-pencil test combined with biodata, an AC, and specific AC exercises combined with two interviews and reference checks. The AC was far superior to the first procedure. The highest utility index was obtained for the third procedure, namely selected AC exercises. Along these lines Payne et al.'s (1992) case study showed how the costs of an AC could be seriously (by almost 25 per cent) reduced. Tziner et al. (1994) investigated the cost-effectiveness of an Israeli AC and found it to be superior to a battery of cognitive tests, even though the validity of the AC was very modest. Recently, Hoffman and Thornton (1997) contrasted a developmental AC and an aptitude test on several criteria such as validity, costs, adverse impact, and utility. Compared to the aptitude test, the developmental AC yielded lower criterion-related validity and was ten times more expensive per candidate. Yet, the AC produced substantially less adverse impact. When cut-off scores were set so as to eliminate adverse impact (as required by the target organisation), this resulted in a higher utility of the AC. In other words, a selection procedure (the AC) with lower validity and higher costs still produced higher utility when considering organisational constraints such as cut-off score choice.

In sum, previous and recent utility research reveals that ACs seem to be superior to more traditional selection techniques (e.g. unstructured interviews, cognitive ability measures, or personality inventories) in terms of selecting high-performing managers. In other words, the benefits in terms of selecting high-performing managers through the use of ACs outweigh the design and administration costs involved.

Assessment centres are fair

According to this slogan there exists neither bias nor potential for unfair discrimination in ACs. Are ACs indeed unbiased or do particular groups (e.g. White, male, and young candidates) systematically receive higher ratings?

Hoffman and Thornton (1997) summarised earlier studies on race bias and concluded that results were evenly split between studies showing no significant differences and studies showing Whites scoring higher on average than other ethnic groups, usually less than one standard deviation. As a possible explanation for these mixed results a recent study (Goldstein et al., 1998) found that the subgroup (Black-White) mean differences were a function of the type of AC exercise employed. Specifically, the subgroup differences varied by the cognitive component of the exercise: ACs exhibited more race bias, if they consisted of a larger number of exercises with a cognitive component (e.g. an in-basket). Along these lines Baron and Janman's (1996) comprehensive treatment of bias at the level of the AC exercise might be insightful. Baron and Janman concluded that with respect to in-baskets possible sources of biases included the presentation and language skills required, inflating women's scores and people from immigrant backgrounds. With respect to group exercises, assessor-assessee effects and group composition effects accounted for small amounts of variance in scores. Baron and Janman (1996) also signalled the dearth of studies regarding other AC exercises such as fact-findings or role-plays.

Other studies focused on assessee gender bias. Again, results were equivocal. In some studies men and women performed equally well (Bobrow and Leonards, 1997; Weijerman and Born, 1995), whereas other studies found evidence for subtle assessee gender bias favouring women (Neubauer, 1990; Schmitt, 1993; Shore, 1992). For instance, Shore (1992) reported that men and women performed equally on overall management potential and interpersonal skills but that women obtained consistently higher ratings on performance-related skills. Regarding the influence of the gender of the assessor, most studies found that ratings of male and female assessors did not differ from each other (Binning et al., 1995; Walsh et al., 1987; Weijerman and Born, 1995). However, Shore et al. (1997) reported that in a role-play female assessors gave significantly higher ratings to both men and women assessees than did male assessors. In two other role-plays no effect of assessor gender was found.

The effects of age bias have also been studied in relation to ACs. Bobrow and Leonards (1997) found very small differences between candidates younger than 40 and candidates 40 and older. However, Clapham and Fulford (1997) reported negative correlations between age and AC ratings after controlling for education, years of service, and gender. In particular, candidates age 40 and over received significantly lower ratings than candidates younger than 40.

In short, despite some evidence for subtle bias, the majority of studies attest to the widely held view that ACs are reasonably unbiased regarding race and gender. Age bias, however, seems to be a more severe problem as younger candidates receive markedly higher ratings.

Candidates react enthusiastically to assessment centres

This slogan looks at ACs from the perspective of the candidate. Research on applicant reactions to ACs has a long research tradition. A first line of research has focused on the reactions of participants after an

AC. Overall, reactions to the AC program and exercises were highly favourable. For instance, Dodd (1977) concluded that a large majority of the participants felt that the AC program was relevant and would recommend it to a friend. Participants responded also very positively to the feedback sessions. Similar findings were obtained in studies conducted in the UK (Dulewicz et al., 1983; Iles and Robertson, 1989), in Canada (Bourgeois et al., 1975), in Germany (Sichler, 1991), and in the Netherlands (Jansen and Stoop, 1995; Zaal, 1990). Apart from these favourable reactions, rejected candidates were typically somewhat less positive (e.g. Teel and DuBois, 1983). In addition, many participants often felt anxious about going through an AC process. Specifically, Fletcher and Kerslake (1992) and Fletcher et al. (1997) found that about 45 per cent of participants reported on stress and anxiety in the post AC questionnaire. However, using established measures of state, trait, and test anxiety Fletcher et al. (1997) did not find support for the relationship between increased anxiety and poor AC performance.

Another line of research has examined applicant reactions to work simulations (which are essential parts of ACs) vis-à-vis other selection devices. For example, Rynes and Connerly (1993) compared 13 selection techniques and concluded that applicants preferred selection methods, which they perceived as job-related. Both written and oral simulations and tests with business-related content scored highest (see also Kluger and Rothstein, 1993). In the UK Iles and Mabey (1993) found that assessment and development centres and interviews were seen as more fair, accurate, and helpful for personal and career development than biodata, personality inventories, or cognitive ability measures.

Recently, studies have applied parts of Gilliland's (1993) justice model of applicant reactions to a selection process. Using this model Kravitz et al. (1996) discovered that students perceived work samples and interviews as the most job-related, fair, and respectful to privacy. In another study students rated ten selection procedures on process favourability and on procedural justice dimensions (Steiner and Gilliland, 1996). Work samples (and interviews) received again the highest overall favourability rating. Additionally, work samples scored highest on the "respectful of privacy" and "perceived face validity" dimensions.

These positive findings were confirmed by two studies, which used samples of real applicants. First, Smither et al. (1993) showed that managers perceived in-baskets, leaderless group discussions, interviews, and cognitive tests with concrete item types as having the greatest predictive and face validity. Second, Macan et al. (1994) examined manufacturing applicants' perceptions of cognitive ability tests and ACs. Applicants reacted favourably to both procedures but viewed the AC as significantly more face valid than the cognitive test battery. After the AC candidates were also more satisfied with the selection process and were more attracted to the organisation than they were previously.

In conclusion, empirical evidence lends support for the slogan that candidates respond enthusiastically to ACs, although low scorers should be given more attention. Apparently, applicants react positively to ACs because centres include work samples and simulation exercises. Research has identified the perceived job relatedness to be the main determinant of the favourable reactions to these simulation-based techniques.

Assessment centres have a potent influence on candidates

Many studies have investigated this slogan and have examined whether the general AC experience, the AC results, and the feedback session have far-reaching consequences for assessees. First, some studies have linked the impact of the AC to variables such as self-perceptions and self-esteem. George and Smith

(1990) reported that self-assessment ratings after the AC were lower than self-assessment ratings prior to the AC. Schmitt et al. (1986) demonstrated that AC participation resulted in changes in self-perceptions of interpersonal and administrative skills. These changes were directly associated with the skills evaluated in the AC. In similar vein, after the AC, candidates were found to be more accurate in their self-assessment (Fletcher and Kerslake, 1992). This increased self-accuracy persisted six months after the feedback. This was not true for unsuccessful candidates, who misjudged the effectiveness of their performance. Another study (Schuler and Fruhner, 1993) focused on changes in the self-concept. For the general self-concept no significant differences were found over time. However, specific facets of the self-concept (e.g. academic self-concept, social self-concept, etc.) were slightly lower during the AC but rose after the feedback session. Schuler and Fruhner (1993) argued that the AC experience and the inherent self-comparison processes led to a lowering of the self-concept, which was subsequently enhanced by the more encouraging feedback.

Other studies have examined how applicant experiences at the AC impact on later career attitudes. For example, Bray et al. (1974) demonstrated that individuals receiving less favourable evaluations after an AC reported less career salience and less upward mobility motivation than people receiving favourable evaluations. Similarly, Slivinski et al. (1979, cited in Noe and Steffy, 1987) found that 79 per cent of the participants reported some change in their career plans as a result of the AC experience. Another study revealed that AC reactions were significantly associated with subsequent job involvement and career exploration behaviour (Noe and Steffy, 1987). Finally, Iles and Robertson (1989) showed that candidates were more likely to think of changing their careers and having less definite career goals after an AC. Nevertheless, they experienced neither change in organisational commitment nor change in job involvement.

A last group of studies have used a variety of variables to determine the impact of the AC experience. For example, Robertson et al. (1991) made comparisons between pass and fail groups on psychological health, self-esteem, organisational commitment, and career withdrawal cognitions. They discovered no significant effects of negative AC results on self-esteem and psychological health. However, negative outcome was found to be associated with reduced organisational commitment and increased career withdrawal cognitions. Fletcher (1991) monitored the long-term effects of an AC at three specific time periods (i.e. before, immediately after, and six months after the AC). Fletcher discovered that self-esteem and job involvement rose immediately after the centre but then dropped back to their pre-assessment levels, whereas work ethic and competitiveness changed significantly in the opposite direction. Job involvement dropped markedly after six months. There were also differences between successful and unsuccessful candidates. The latter ones exhibited significant reductions in self-esteem, the work ethic component of need for achievement, and job involvement, six months after the centre. Recently, Francis-Smythe and Smith (1997) confirmed that the assessment outcome affected post-assessment psychological characteristics and work-related attitudes. In particular, the effect of AC outcomes on psychological variables seemed to work through the individual's perception of the career impact of the AC. This perception was well predicted by candidates' perceived quality of feedback and current career position. Good feedback helped careers substantially, and the nearer the person's current position was to the role for which they were being assessed the more likely they were to benefit from the process. Despite the majority of the aforementioned studies reported some impact validity for ACs, Williams and Harrington (1998), concluded that their AC exerted virtually no impact on a wide array of psychological variables. They also noted that in prior research a variety of variables had been measured by many different questionnaires, impairing generalizability.

In short, most studies report that participating in an AC in general and a positive AC outcome in particular positively influence self-perceptions and later career exploration behaviour, providing support for the slogan. However, research also shows that attention should be paid to the effects of a negative decision after attending an AC and to good feedback.

As a candidate it is difficult to prepare oneself for an assessment centre

This slogan pertains to the presumed lack of effects of coaching to improve candidates' subsequent AC performance. To test this slogan we found research studies on whether coaching affected performance in in-baskets, role-plays, and group discussions.

Regarding in-basket exercises, three studies discovered that students were able to improve their performance on specific aspects of an in-basket after self-study of workbooks designed to help them structure their thinking about the task (Brannick et al., 1989; Brostoff and Meyer, 1984; Gill, 1982). In addition, these workbooks advised the assessees to search for connections among in-basket items and provided coaching on organising and planning, time management, and delegation.

Moses and Ritchie (1976) examined the effects of a supervisory relationship training on role-play performance. First-level supervisors, who had followed this behaviour modelling program, were significantly more effective in dealing with problem subordinates in three role-plays than untrained supervisors.

With respect to group discussion exercises, Petty (1974) found that even a 15 minute briefing improved students' group discussion performance. Similar evidence comes from an unpublished study by Bracken (1989). He found that a training session improved both leaderless group discussion knowledge and leaderless group discussion performance significantly over mere leaderless group discussion participation or no-training. Kurecka et al. (1982) reported that a group of students, who participated in a leaderless group discussion and received behaviourally specific feedback on their performance, did better in a subsequent leaderless group discussion than a control group and a group, who received bad advice on how to behave in a group exercise.

Recently, MacKinnon-Slaney (2000) studied the impact of candidate tutorial programs. Candidates, who attended a tutorial program, performed significantly better in written AC exercises compared to candidates who did not attend a tutorial program. No differences were found for performance in oral AC exercises. In this study candidate tutorials described the content of the AC exercises, the evaluation methods, and provided candidates with suggestions as to how they can study and practice for the exercises.

Finally, Dulewicz and Fletcher (1982) did not focus on assessee coaching but investigated whether previous experience of specific managerial activities relevant to AC exercises significantly affected AC performance. In general, no such effects were found. Only in the business plan presentation exercise a significant effect for prior experience in business planning (and not in making formal presentations) was found.

In sum, research on the trainability of AC performance does not lend support for the slogan that it is difficult to prepare oneself for an AC. Assessee coaching appears to significantly improve performance in in-baskets, role-plays, and group discussions. However, due to the limited number of studies, no definite

conclusions can be drawn. Note also that only a snapshot of possible coaching tactics have been studied and the effects of assessee coaching on subsequent job performance are unknown.

Assessment centres provide insight into candidates' personality

Delivering presentations, leading group discussions, or coaching disgruntled subordinates are only a few examples of popular AC exercises. Hence, brochures promoting ACs often state that AC exercises are ideal instruments to obtain insightful information of candidates' personality. Researchers have tackled this slogan by relating AC scores (i.e. overall assessment rating, final dimensional scores, etc.) to an array of external measures such as cognitive ability measures or personality inventories. Scholz and Schuler's (1993) meta-analysis included 51 such studies. Intelligence correlated 0.33 with the overall assessment rating, which increased to 0.43 when corrected for unreliability. Besides intelligence the overall AC rating also tended to correlate 0.23 with dominance, 0.30 with achievement motivation, 0.31 with social competence, and 0.26 with self-confidence (all were corrected for unreliability). Fleenor (1996) found that the personality trait "exhibition" was significantly correlated with all ten AC dimensions measured. Apparently, participants who were "good actors" and highly competitive were rated significantly higher in the AC. Fleenor (1996) also concluded that the trait "aggression" significantly correlated with seven AC dimensions, and the trait "dominance" with five dimensions. Waldman and Atwater (2000) uncovered that people high on rugged individualism obtain higher ratings in the AC. In another recent study Furnham et al. (1997) validated the NEO Personality Inventory using assessors' ratings. A clear pattern emerged with conscientiousness and extraversion having strongest and most frequent correlations. Contrary to these statements, Goffin et al. (1996) reported a marked lack of substantial correlations between personality and AC scores. They also found significant incremental validity for both personality and dimensional AC scores over one another.

In sum, studies on the overlap between personality constructs and ACs show a somewhat equivocal picture. Whereas some studies report significant correlations between AC scores and dimensions such as interpersonal competence, dominance, rugged individualism, exhibition or achievement motivation, other studies conclude that "personality and ACs sample different domains which in turn predict relatively different aspects of job performance (Goffin et al., 1996, p. 753)".

A possible explanation for these ambiguous results may be that final dimensional ratings and overall assessment ratings are too broad to be effective in this comparison. Therefore, an interesting avenue for future studies may consist in relating ratings on personality inventories to dimensional ratings in separate AC exercises. For instance, a recent study (Spector et al., 2000) examined cognitive ability and personality correlates of AC exercise ratings. Ratings of interpersonal exercises correlated with interpersonal personality constructs (emotional stability, extraversion, and openness to experience) and ratings of cognitive problem solving exercises correlated with cognitive ability and conscientiousness.

Conclusions

This paper compared eight common beliefs about ACs to the extant empirical evidence. Two slogans (i.e. "ACs are expensive" and "As a candidate it is difficult to prepare oneself for an AC") were not supported by empirical research. In particular, earlier and recent studies showed that ACs were superior to unstructured interviews, measures of cognitive ability, and personality inventories in terms of utility, and that coaching effects were observed in in-baskets, role-plays, and group discussions. We could not draw

definite conclusions regarding the slogan that "ACs provide insight into candidates' personality" because the research results currently available were equivocal. We found partial support for the remaining five slogans (i.e. "Trained assessors speak the same language", "ACs work: they show high predictive validity", "ACs are fair", "Candidates react enthusiastically to ACs", and "ACs have a potent influence on candidates"). In fact, a thread running through these slogans was that they were true for some centres but not for others as various moderating factors related to AC design played an important role. For example, the type of training given to assessors or the inclusion of an integrative discussion were among others identified as important factors to enhance inter-rater reliability. In similar vein, various factors (e.g. number of exercises, type of assessor, the degree of assessee-assessor acquaintance) substantially moderated AC validity. The same argument was true for subgroup bias. Generally, ACs exhibited only subtle race or gender bias effects. However, if ACs contained many cognitive-oriented exercises (e.g. in-baskets), more race bias was found. Research studies also attested to the enthusiastic reaction of candidates to ACs in general and to the simulations used in particular. Moreover, some studies reported that participating as an assessee positively influenced later career exploration behaviour and psychological characteristics such as better self-perceptions. Yet, attention should be paid to less successful candidates. Again, a design factor such as the provision of good feedback was found to influence these reactions.

Where should AC research go from here? In order to advance the field we believe that studies with a broad research design should be conducted. There have already been many separate studies on criterion-related validity, applicant reactions, fairness, etc. What we need now are studies, which combine multiple research perspectives or broaden existing research perspectives. Examples of integrative studies are studies, which address both adverse impact and utility (e.g. Hoffman and Thornton, 1997) or which combine applicant reactions and subgroup differences. An example of broadening existing perspectives is the inclusion of extra-role behaviours (e.g. volunteering to help colleagues, skipping breaks, etc.) or organisational criteria (see Russell and Domm, 1995) in the criterion measures of validity studies. Research on reliability should also be broadened and address other forms of reliability (e.g. test-retest, alternate form reliability, etc.). To this end, generalizability analysis (Cronbach et al., 1972; Marcoulides, 1989) might serve as a fruitful framework in AC research. As opposed to classical test theory, generalizability theory regards measurement error as multifaceted. Accordingly, it permits the simultaneous estimation of different sources of variance (e.g. assessors, exercises, dimensions, sessions, administrators, role-players, etc.) inherent in ratings.

We also encourage that future studies examine the AC process. As argued by Lievens and Klimoski (2001) we particularly need studies which provide a better understanding of the individual and collective processes that affect the quality of assessor decisions. These studies could among others help to explain why ACs work. In addition, it is important to audit the whole AC design process for gender, race, or age bias instead of concentrating only on the AC ratings. In particular, areas for future research include procedures related to the selection of AC dimensions, the nomination of AC candidates, the integrative assessor discussion, the use of role-players, or the feedback to candidates (see Baron and Janman, 1996; Blanksby and Iles, 1995). At the same time applicant reactions to these procedures can be investigated.

Finally, we make a plea that future studies try to validate new applications of ACs. For example, there is no validity research about ACs for international selection or team selection. There is also no research, which contrasts the incremental validity or utility of ACs to multimedia tests or video-based situational judgement tests.

References

1. Abrahamson, E. (1996), “Management fashion”, Academy of Management Review, Vol. 21, pp. 254-85.

2. Baron, H. and Janman, K. (1996), “Fairness in the assessment centre”, International Review of Industrial and Organizational Psychology, Vol. 11, pp. 61-114.

3. Bartels, L.K. and Doverspike, D. (1997), “Effects of disaggregation on managerial assessment center validity”, Journal of Business and Psychology, Vol. 12, pp. 45-53.

4. Binning, J.F., Adorno, A.J. and Williams, K.B. (1995, May), “Gender and race effects on behavior checklist and judgmental assessment center evaluations”, paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

5. Blanksby, M. and Iles, P. (1995), “Recent developments in assessment centre theory, practice and operation”, Personnel Review, Vol. 19, pp. 33-44.

6. Bobrow, W. and Leonards, J.S. (1997), “Development and validation of an assessment center during organizational change”, Journal of Social Behavior and Personality, Vol. 12, pp. 217-36.

7. Borman, W.C. (1982), ‘‘Validity of behavioral assessment for predicting military recruiter performance”, Journal of Applied Psychology, Vol. 67, pp. 3-9.

8. Bourgeois, R.P., Leim, M.A., Slivinski, L.W. and Grant, K.W. (1975), “Evaluation of an assessment center in terms of acceptability”, Canadian Personnel and Industrial Relations Journal, Vol. 22, pp. 17-20.

9. Boyle, S., Fullerton, J. and Wood, R. (1995), “Do assessment/development centres use optimum evaluation procedures? A survey of practice in UK organizations”, International Journal of Selection and Assessment, Vol. 3, pp. 132-40.

10. Bracken, D.T. Jr (1989), “Assessment center exercise performance: an examination of the trainability of the leaderless group discussion”, unpublished doctoral dissertation, University of Georgia, Athens, GA.

11. Brannick, M.T., Michaels, C.E. and Baker, D.P. (1989), “Construct validity of in-basket scores”, Journal of Applied Psychology, Vol. 74, pp. 957-63.

12. Bray, D.W., Campbell, R.J. and Grant, D.L. (1974), Formative Years in Business: A Long-term AT&T Study of Managerial Lives, Wiley, New York, NY.

13. Brostoff, M. and Meyer, H.H. (1984), “The effects of coaching on in-basket performance”, Journal of Assessment Center Technology, Vol. 7, pp. 17-21.

14. Burke, M.J. and Frederick, J.T. (1986), “A comparison of economic utility estimates for alternative SD-sub(y) estimation procedures”, Journal of Applied Psychology, Vol. 71, pp. 334-9.

15. Cascio, W.F. and Ramos, R.A. (1986), “Development and application of a new method for assessing job performance in behavioral/economic terms”, Journal of Applied Psychology, Vol. 71, pp. 20-28.

16. Cascio, W.F. and Silbey, V. (1979), “Utility of the assessment center as a selection device’’, Journal of Applied Psychology, Vol. 64, pp. 107-18.

17. Chan, D. (1996), “Criterion and construct validation of an assessment centre”, Journal of Occupational and Organisational Psychology, Vol. 69, pp. 167-81.

18. Clapham, M.M. and Fulford, M.D. (1997), “Age bias in assessment center ratings”, Journal of Managerial Issues, Vol. 9, pp. 373-87.

19. Cook, M. (1998), Personnel Selection, Adding Value through People, John Wiley, Chicester.

20. Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. (1972), The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles, Wiley, New York, NY.

21. Dodd, W.E. (1977), “Attitudes toward assessment center programs”, in Moses, J.L. and Byham W.C. (Eds), Applying the Assessment Center Method, Pergamon Press, New York, NY, pp. 161-84.

22. Dulewicz, V. and Fletcher, C. (1982), “The relationship between previous experience, intelligence and background characteristics of participants and their performance in an assessment centre”, Journal of Occupational Psychology, Vol. 55, pp. 197-207.

23. Dulewicz, V., Fletcher, C. and Wood, P. (1983), “A study of the internal validity of an assessment centre and of participant’s background characteristics and attitudes: a comparison of British and American findings”, Journal of Assessment Center Technology, Vol. 6, pp. 15-24.

24. Fleenor, J.W. (1996), “Constructs and developmental assessment centers: further troubling empirical findings”, Journal of Business and Psychology, Vol. 10, pp. 319-33.

25. Fletcher, C. (1991), “Candidates’ reactions to assessment centres and their outcomes: a longitudinal study”, Journal of Occupational Psychology, Vol. 64, pp. 117-27.

26. Fletcher, C. and Kerslake, C. (1992), ‘‘The impact of assessment centers and their outcomes on participants’ self-assessments”, Human Relations, Vol. 45, pp. 281-9.

27. Fletcher, C., Lovatt C. and Baldry, C. (1997), “A study of state, trait, and test anxiety, and their relationship to assessment center performance”, Journal of Social Behavior and Personality, Vol. 12, pp. 205-14.

28. Francis-Smythe, J. and Smith, P.M. (1997), “The psychological impact of assssment in a development center”, Human Relations, Vol. 50, pp. 149-67.

29. Furnham, A., Crump, J. and Whelan, J. (1997), “Validating the NEO personality inventroy using assessor’s ratings”, Personality and Individual Differences, Vol. 22, pp. 669-75.

30. Gatewood, R., Thornton, G.C. III, and Henessey, H.W. Jr (1990), “Reliability of expertise ratings in the leaderless group discussion”, Journal of Occupational Psychology, Vol. 63, pp. 331-42.

31. Gaugler, B.B., Rosenthal, D.B., Thornton, G.C. and Bentson, C. (1987), “Meta-analysis of assessment center validity”, Journal of Applied Psychology, Vol. 72, pp. 493-511.

32. George, D.I. and Smith, M.C. (1990), “An empirical comparison of self-assessment and organizational assessment in personnel selection”, Public Personnel Management, Vol. 19, pp. 175-90.

33. Gill, R.W.T. (1982, “A trainability concept for management potential and an empirical study of its relationship with intelligence for two managerial skills”, Journal of Occupational Psychology, Vol. 52, pp. 185-97.

34. Gilliland, S.W. (1993), “The perceived fairness of selection systems: an organizational justice perspective”, Academy of Management Review, Vol. 18, pp. 694-734.

35. Goffin, R.D., Rothstein, M.G. and Johnston, N.G. (1996), “Personality testing and the assessment center: incremental validity for managerial selection”, Journal of Applied Psychology, Vol. 81, pp. 746-56.

36. Goldsmith, R.F. (1990), “Utility analysis and its application to the study of the cost effectiveness of the assessment center method”, in Murphy, K.R. and Saal, F.E. (Eds), Psychology in Organizations: Integrating Science and Practice, Erlbaum, Hillsdale, NJ, pp. 95-110.

37. Goldstein, H.W. Yusko, K.P., Braverman, E.P. Smith, D.B. and Chung, B. (1998), “The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises”, Personnel Psychology, Vol. 51, pp. 357-74.

38. Hoffman, C.C. and Thornton, G.C. III (1997), “Examining selection utility where competing predictors differ in adverse impact”, Personnel Psychology, Vol. 50, pp. 455-70.

39. Hogan, J. and Zenke, L.L. (1986), “Dollar-value utility of alternative procedures for selecting school principals”, Educational and Psychological Measurement, Vol. 46, pp. 935-45.

40. Iles, P.A. and Mabey, C. (1993), “Managerial career development techniques: effectiveness, acceptability and availability”, British Journal of Management, Vol. 4, pp. 103-18.

41. Iles, P.A. and Robertson, I.T. (1989), “The impact of personnel selection procedures on candidates”, in Herriot, P. (Ed.), Assessment and Selection in Organizations, Wiley, Chichester.

42. Jansen, P.G.W. and Stoop, B. (1994), “Assessment center graduate selection: decision processes, validity, and evaluation by candidates”, International Journal of Selection and Assessment, Vol. 2, pp. 193-208.

43. Johns, G. (1993), “Constraints on the adoption of psychology-based personnel practices: lessons from organizational innovation”, Personnel Psychology, Vol. 46, pp. 569-92.

44. Jones, A. (1981), “Inter-rater reliability in the assessment of group exercises at the UK assessment centre”, Journal of Occupational and Organisational Psychology, Vol. 54, pp. 79-86.

45. Jones, A., Herriot, P., Long, B. and Drakeley, R.H. (1991), “Attempting to improve the validity of a well-established assessment centre”, Journal of Occupational Psychology, Vol. 64, pp. 1-21.

46. Keenan, T. (1995), “Graduate recruitment in Britain: a survey of selection methods used by organizations’’, Journal of Organizational Behavior, Vol. 16, pp. 303-17.

47. Kluger, A.N. and Rothstein, H.R. (1993), “The influence of selection test type on applicant reactions to employment testing”, Journal of Business and Psychology, Vol. 8, pp. 3-25.

48. Kravitz, D.A., Stinson, V. and Chavez, T.L. (1996), “Evaluations of tests used for making selection and promotion decisions”, International Journal of Selection and Assessment, Vol. 4, pp. 24-34.

49. Kurecka, P.M., Austin, J.M., Johnson, W. and Mendoza, J.L. (1982), “Full and errant coaching effects on assigned role leaderless group discussion performance”, Personnel Psychology, Vol. 35, pp. 805-12.

50. Lievens, F. (2001), “Assessor training strategies and their effects on accuracy, inter-rater reliability, and discriminant validity’’, Journal of Applied Psychology, Vol. 86.

51. Lievens, F. (in press), “Assessors and use of assessment center dimensions: a fresh look at a troubling issue”, Journal of Organizational Behavior.

52. Lievens, F. and Goemaere, H. (1999), “A different look at assessment centers: views of assessment center users’’, International Journal of Selection and Assessment, Vol. 7, pp. 215-9.

53. Lievens, F. and Klimoski, R.J. (2001), “Understanding the assessment center process: where are we now?”, in Cooper, C.L. and Robertson, I.T. (Eds), International Review of Industrial and Organizational Psychology, Vol. 16, John Wiley, Chichester, pp. 245-86.

54. Macan, T.H., Avedon, M.J., Paese, M. and Smith, D.E. (1994), “The effects of applicants’ reactions to cognitive ability tests and an assessment center”, Personnel Psychology, Vol. 47, pp. 715-38.

55. MacKinnon-Slaney, I.M. (2000), “Candidate tutorials: do they improve candidate performance?”, paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

56. Marcoulides, G.A. (1989), “The application of generalizability analysis to observational studies”, Quality and Quantity, Vol. 23, pp. 115-27.

57. Moser, K., Schuler, H. and Funke, U. (1999), “The moderating effect of raters’ opportunities to observe ratees’ job performance on the validity of an assessment centre”, International Journal of Selection and Assessment, Vol. 7, pp. 133-41.

58. Moses, J.L. and Ritchie, R.J. (1976), “Supervisory relationships training: a behavioral evaluation of a behavior modeling program”, Personnel Psychology, Vol. 29, pp. 337-43.

59. Neubauer, R. (1990), “Women in the career assessment center – a victory?”, (in German), Zeitschrift für Arbeits und Organisationspsychologie, Vol. 34, pp. 29-36.

60. Noe, R.A. and Steffy, B.D. (1987), “The influence of individual characteristics and assessment center evaluation on career exploration behavior and job involvement”, Journal of Vocational Behavior, Vol. 30, pp. 187-202.

61. Payne, T., Anderson, N.R. and Smith, T. (1992), ‘‘Assessment centers, selection systems and cost effectiveness: an evaluative case study”, Personnel Review, Vol. 21, pp. 48-56.

62. Petty, M.M. (1974), “A multivariate analysis of the effects of experience and training upon performance in a leaderless group discussion”, Personnel Psychology, Vol. 27, pp. 271-82.

63. Robertson, I.T., Iles, P.A., Gratton, L. and Sharpley, D. (1991), “The impact of personnel selection and assessment methods on candidates”, Human Relations, Vol. 44, pp. 963-82.

64. Russell, C.J. and Domm, D.R. (1995), “Two field tests of an explanation of assessment centre validity”, Journal of Occupational and Organizational Psychology, Vol. 68, pp. 25-47.

65. Rynes, S.L. and Connerly, M.L. (1993, “Applicant reactions to alternative selection procedures”, Journal of Business and Psychology, Vol. 7, pp. 261-78.

66. Schmitt, N. (1993), “Group composition, gender, and race effects on assessment center ratings”, in Schuler, H., Farr, J.L. and Smith M. (Eds), Personnel Selection and Assessment: Individual and Organizational Perspectives, Lawrence Erlbaum, Hillsdale, NJ, pp. 315-32.

67. Schmitt, N., Ford, J.K. and Stults, D.M. (1986), “Changes in self-perceived ability as a function of performance in an assessment centre”, Journal of Occupational Psychology, Vol. 59, pp. 327-35

68. Scholz, G. and Schuler, H. (1993), “The nomological network of the assessment center: a metaanalyses”, (in German), Zeitschrift für Arbeits- und Organisationspsychologie, Vol. 37, pp. 73-85.

69. Schuler, H. and Fruhner, R. (1993), “Effects of assessment center participation on self-esteem and on evaluation of the selection situation”, in Schuler H., Farr, J.L. and Smith, M. (Eds), Personnel Selection and Assessment: Individual and Organizational Perspectives, Lawrence Erlbaum, Hillsdale, NJ, pp. 109-24.

70. Shackleton, V. and Newell, S. (1994), “European management selection methods: a comparison of five countries”, International Journal of Selection and Assessment, Vol. 2, pp. 91-102.

71. Shore, T.H. (1992), “Subtle gender bias in the assessment of managerial potential”, Sex Roles, Vol. 27, pp. 499-515.

72. Shore, T.H., Taschian, A. and Adams, J.S. (1997), “The role of gender in a developmental assessment center”, Journal of Social Behavior and Personality, Vol. 12, pp. 191-203.

73. Sichler, R. (1991), ‘‘Experiences and activities with an assessment center procedure: an empirical contribution to the ‘social validity’ of test situations”, (in German), Zeitschrift für Arbeits- und Organisationspsychologie, Vol. 33, pp. 139-45.

74. Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K. and Stoffey, R.W. (1993), “Applicant reactions to selection procedures”, Personnel Psychology, Vol. 46, pp. 49-78.

75. Spector, P.E., Schneider, J.R., Vance, C.A. and Hezlett, S.A. (2000), “The relation of cognitive ability and personality traits to assessment center performance”, Journal of Applied Social Psychology, Vol. 30, pp. 1474-91.

76. Steiner, D.D. and Gilliland, S.W. (1996), “Fairness reactions to personnel selection techniques in France and the United States”, Journal of Applied Psychology, Vol. 81, pp. 134-47.

77. Teel, K. and DuBois, H. (1983), “Participants’ reactions to assessment centers”, Personnel Administrator, Vol. 28, pp. 85-91.

78. Tziner, A. and Dolan, S. (1982), “Validity of an assessment center for identifying future female officers in the military”, Journal of Applied Psychology, Vol. 67, pp. 728-36.

79. Tziner, A., Ronen, S. and Hacohen, D. (1993), “A four year validation study of an assessment center in a financial corporation”, Journal of Organizational Behavior, Vol. 14, pp. 225-37.

80. Tziner, A., Meir, E.I., Dahan, M. and Birati, A. (1994), “An investigation of the predictive validity and economic utility of the assessment center for the high-management level”, Canadian Journal of Behavioural Science, Vol. 26, pp. 228-45.

81. Waldman, D.A. and Atwater, L.E. (2000), “The role of rugged individualism in assessment center performance: does John Wayne make a good candidate?”, paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

82. Walsh, J.P., Weinberg, R.M. and Fairfield, M.L. (1987), “The effects of gender on assessment centre evaluations’’, Journal of Occupational Psychology, Vol. 60, pp. 305-9.

83. Weijerman, E.A.P. and Born, M.P. (1995), “The relationship between gender and assessment center scores”, (in Dutch), Gedrag en Organisatie, Vol. 8, pp. 284-92.

84. Williams, R.S. and Harrington, E. (1998, August), “Assessment centre impact validity: when no impact is a good thing”, paper presented at the International Congress of Applied Psychology, San Francisco, CA.

85. Zaal, J.N. (1990), “Assessment centers at the Dutch central government”, in Hambleton, R.K. and Zaal, J.N. (Eds), Advances in Educational and Psychological Testing: Theory and Applications, Kluwer, Boston, pp. 227-50.